Statistics Assignment

Submitted by: Sachin Katiyar

The pharmaceutical company Sun Pharma is manufacturing a new batch of painkiller drugs, which are due for testing. Around 80,000 new products are created and need to be tested for their time of effect (which is measured as the time taken for the drug to completely cure the pain), as well as the quality assurance (which tells you whether the drug was able to do a satisfactory job or not).

Question-1:

The quality assurance checks on the previous batches of drugs found that — it is 4 times more likely that a drug is able to produce a satisfactory result than not.

Given a small sample of 10 drugs, you are required to find the theoretical probability that at most, 3 drugs are not able to do a satisfactory job.

a) Propose the type of probability distribution that would accurately portray the above scenario, and list out the three conditions that this distribution follows.

b) Calculate the required probability.

Answer:

Here, Random variable X = No. of drugs that are not able to do a satisfactory job.

Let the probability that a drug will not produce a satisfactory result is p.

Since a drug is able to produce a satisfactory result is 4 times more likely than not. Therefore, probability that a drug will produce a satisfactory result is 4p. Since probability of all outcome sums up to 1, therefore:

- ⇨ $p+4p = 1$
- ⇨ $5p = 1$
- ⇨ $p = (1/5) = 0.2$

Thus, probability that a drug will not produce a satisfactory result is $p = 0.2$

Here, Total no of drugs in given sample i.e. $n = 10$

We have to find theoretical probability that at most 3 drugs (i.e. $x=3$) are not able to do a satisfactory job i.e. the cumulative probability of the random variable X, taking a value lesser than or equal to $x=3$

Mathematically speaking:

$P(X \leq 3) = F(3) = ?$ (To be find)

a) Since given scenario follows below three conditions, i.e.
    - The total number of trails is fixed, $n = 10$

- Each trial is binary i.e. having only two possible outcome, success and failure. Based on the question asked, success scenario is that a drug will not produce a satisfactory result while the failure scenario is that a drug will produce a satisfactory result.
- The probability of success is same for all the trails i.e. probability that a drug will not produce a satisfactory result is 0.2

Therefore, given scenario follows the Binomial Distribution, which is given by:

$$P(X=r) = {}^{n}C_{r}\,(p)^{r}\,(1-p)^{n-r}$$

Where:

n is the number of trails

r is the no of success after n trails

p is the probability of success in one trial

But we have to find that at most 3 drugs are not able to do a satisfactory job which is given by:

$$F(x) = F(3) = P(X<=x) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

Where:

F(x) = Cumulative Probability of random variable X taking a value less than x

b) Calculating the required probability:

$$F(3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$$

Here n=10, p = 0.2

$$P(X=r) = {}^{n}C_{r}\,(p)^{r}\,(1-p)^{n-r}$$

$$P(X=0) = {}^{10}C_{0}\,(0.2)^{0}\,(1-0.2)^{10-0}$$

$$= 1 \times 1 \times 0.8^{10}$$

$$= 0.1073741824$$

$$P(X=1) = {}^{10}C_{1}\,(0.2)^{1}\,(1-0.2)^{10-1}$$

$$= 10 \times 0.2 \times 0.8^{9}$$

$$= 0.268435456$$

$$P(X=2) = {}^{10}C_{2}\,(0.2)^{2}\,(1-0.2)^{10-2}$$

$$= 45 \times 0.04 \times 0.8^{8}$$

$$= 0.301989888$$

$$P(X=3) = {}^{10}C_{3}\,(0.2)^{3}\,(1-0.2)^{10-3}$$

$$= 120 \times 0.008 \times 0.8^{7}$$

= 0.201326592

F (3) = 0.1073741824 + 0.268435456 + 0.301989888 + 0.201326592

= 0.8791261184

= 0.8791 (rounding off)

Therefore, probability of at most 3 drugs are not able to do a satisfactory job is 0.8791

Question-2:

For the effectiveness test, a sample of 100 drugs was taken. The mean time of effect was 207 seconds, with the standard deviation coming to 65 seconds. Using this information, you are required to estimate the range in which the population mean might lie — with a 95% confidence level.

a) Discuss the main methodology using which you will approach this problem. State all the properties of the required method. Limit your answer to 150 words.

b) Find the required range.

Answer:

a) Since it is difficult to calculate mean of all 80000 products because of time and/or money constraints, it will be beneficial to find the mean for only a small representative sample. Therefore, we will approach towards a theorem called **Central Limit Theorem** which states that
   - Given a sufficiently large sample size from a population, the mean of all samples from the same population will be approximately equal to the mean of the population.
   - As the sample sizes get larger (n > 30), the distribution of means calculated from repeated sampling will approach normality.
   - Sample Distribution's Standard Deviation (Sample Error) = Population Standard Deviation / sqrt (Sample Size)

b) Here, size of sample given: n = 100
   Sample Mean $\bar{X}$ = 207 seconds
   Standard Deviation S = 65 seconds
   Confidence Level y% = 95%

   Z-score associated with a 95% confidence level = 1.96

   Margin of error = $\frac{z^* s}{\sqrt{n}}$ = (1.96 x 65) / sqrt (100) = 12.74

   Calculating confidence interval = $(\bar{X} - \frac{z^* s}{\sqrt{n}}, \bar{X} + \frac{z^* s}{\sqrt{n}})$

   = (207 - 12.74, 207 + 12.74)

   = (194.26, 219.74)

   Therefore, Population mean might lie between range (194.26, 219.74)

Question 3:

a) The painkiller drug needs to have a time of effect of at most 200 seconds to be considered as having done a satisfactory job. Given the same sample data (size, mean, and standard deviation) of the previous question, test the claim that the newer batch produces a satisfactory result and passes the quality assurance test. Utilize 2 hypothesis testing methods to make your decision. Take the significance level at 5 %. Clearly specify the hypotheses, the calculated test statistics, and the final decision that should be made for each method.

b) You know that two types of errors can occur during hypothesis testing — namely Type-I and Type-II errors — whose probabilities are denoted by α and β respectively. For the current sample conditions (sample size, mean, and standard deviation), the value of α and β come out to be 0.05 and 0.45 respectively.

Now, a different sampling procedure (with different sample size, mean, and standard deviation) is proposed so that when the same hypothesis test is conducted, the values of α and β are controlled at 0.15 each. Explain under what conditions would either method be more preferred than the other, i.e. give an example of a situation where conducting a hypothesis test having α and β as 0.05 and 0.45 respectively would be preferred over having them both at 0.15. Similarly, give an example for the reverse scenario - a situation where conducting the hypothesis test with both α and β values fixed at 0.15 would be preferred over having them at 0.05 and 0.45 respectively. Also, provide suitable reasons for your choice (Assume that only the values of α and β as mentioned above are provided to you and no other information is available).

Answer:

a) Formulating Hypotheses:
- Null Hypothesis H_0 : Time of effect for a painkiller drug ≤ 200 seconds
- Alternate Hypothesis H_1 : Time of effect for a painkiller drug > 200 seconds

To check if there is enough evidence to conclude if the hypotheses about time of effect for a painkiller is true or not, we will follow below two methods one by one:

1. Critical Value Method
2. p - value Method

Critical Value Method

- This is a one-tailed test. Based on alternate hypothesis, we can say that we will have only one critical region on the right side with a total area of 0.05 (since significance level is given as α = 5%).
- This means that the area till the critical point will be 1 - 0.050 = 0.950. i.e. Area of acceptance region will be 0.95 and area of rejection region will be 0.05 .
- Now, calculating the z-value of 0.950 from the z-table that comes out to be 1.645 (Since z value of 0.9505 given in table is 1.65 and z value of 0.9495 given in table is 1.64, so taking the average of these two, the z-score for 0.950 is 1.645)
- Since critical region lies on the right side, we are moving forward to calculate Upper Critical Value (UCV).

  Here, μ = 200 seconds

Z_c = 1.645
σ = 65 seconds
n = 100

UCV = $\mu + (Z_c \times \frac{\sigma}{\sqrt{n}})$

UCV = $200 + (1.645 \times (\frac{65}{\sqrt{100}}))$
= 200 + (1.645 x 6.5)
= 210.6925

Since sample mean (207 seconds) is less than UCV (210.6925 seconds) therefore sample mean lies in the acceptance region.

Decision: Fail to reject the null hypothesis

p - value Method

- Calculating z score for the sample mean point $\bar{x}$ = 207 seconds using below formula:
  z score = $\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{N}}}$
  z score = (207 – 200) / (65/sqrt (100))
  z score = 7/6.5
  z score = 1.076
- Calculating the p – value from the cumulative probability for z score =1.076 using z table:
  The value corresponds to 1.076 in z table is 0.8590.
  Since z score is positive therefore sample mean will be on the right side of the distribution mean. Also, it is one tail test. Therefore p-value would be (1 – 0.8590)
  p – value = 0.1410

Since p – value (0.1410) > significance value (α = 0.05)

Decision: Fail to reject the null hypothesis

b) Type I error (α) : Null hypothesis is true but rejected.
Type II error (β) : Fail to reject the false null hypothesis.

⇨ Below is the example where type I error (as 0.05) and type II error (as 0.45) is preferred over type I and type II error (as 0.15).

Place A and Place B is separated by a lake that has crocodiles. You have to go to place B from place A. You can only reach place B from place A by swimming through that lake.

H_0 : Lake has crocodiles.

H_1 : Lake do not have crocodiles.

- Here α cannot be large, as if we reject H_0 which is true, we will be eaten by crocodiles. In this case, α must be as less as 0.05
- β can be large as 0.45 since H_0 is false and we failed to reject the false null hypothesis, we are not in any harm. Since H_0 is false which suggests that there are crocodile in lake is a false information and we failed to reject this information thus we can easily go from place A to place B without any harm.

⇨ Below is the example where type I and type II error (as 0.15) is preferred over type I error (as 0.05) and type II error (as 0.45)

A new X app is launched which is as secure as Google app. Both does the same functionality. You have already installed Google app in your phone but your friend asked to install new X app as this will help him in earning referral money. A news has come saying that some newly launched apps are stealing contact information from your phone while company that owns X app rejects this news for their app. Your have to take decision to install or not install the X app.

H_0 : App X is safe and secure to use.
H_1 : App X is not safe and secure to use.

- Here α can be large as 0.15 since this rejects H_0 which is true. This will not do any harm to you as you can continue to use Google app instead X app since we reject that app X is safe and secure to use which is correct.
- β cannot be as large as 0.45, it must be small as 0.15. Since H_0 is false which suggests that App X is safe and secure is a false information and we failed to reject this information which can cause harm to us. App X can steal our contact information and can misuse it.

Question-4:

Now, once the batch has passed all the quality tests and is ready to be launched in the market, the marketing team needs to plan an effective online ad campaign to attract new customers. Two taglines were proposed for the campaign, and the team is currently divided on which option to use.

Explain why and how A/B testing can be used to decide which option is more effective. Give a stepwise procedure for the test that needs to be conducted.

Answer:

A/B Testing is an online marketing tactic to check which version of product does user prefer. This is basically a way to check how a variable affects the audience reaction.

<u>How A/B testing works:</u>

- A/B testing is performed on two different versions of products, let it be webpage.
- Original version is called as control while newer version which has to be launched is called variation.

- Control is shown to half of the public while variation is shown to half of the public who are browsing that web page.
- After that, involvement of users with each change is measured and analysed. Based on that company takes decision whether new version is good/bad than the original version or it is all same.

Why to do A/B test:

- A/B test allows companies and individuals to make user experience that suits best to their users based on their involvement.
- User interaction with each change done on newer version is noticed and based on that decisions are taken.
- Also, it helps market strategy team to know which version attract more users.
- Sometime it rejects user experience that was thought to be the best to achieve a particular goal as reality can be different than the theory.

Steps for conducting A/B testing:

Below are the steps to perform A/B testing –

- Collecting data for analysis will be the first task. For this we have to know the high traffic areas to get data as quickly as possible.
- Identifying goals that you have to check on, i.e. whether new or old version is good. It can be anything (button shape, text written on button, payment interface etc)
- Generating A/B testing ideas and hypotheses for why new version is better than the old one. Also arrange them in terms of expected impact.
- Now create variation on your website using A/B testing software like changing color or orientation of tab etc.
- After that make your experiment live and wait for the user to interact. Website versions will be allocated randomly to each user. Parallelly analyse each interaction of user with each variable you have changed.
- Once it is complete, analyse the results. A/B testing software will present you with all the data and will show you how both versions have impacted users to attract their attention.